

AI in Retinal Imaging: How Ophthalmologists Handle Trust and Accountability

A global peer report on retinal AI utility, clinical disagreement, image-quality failure, autonomous screening, automation bias, explainability, workflow barriers, and accountability.

Ophthalmologists · 5 countries · Global peer perspectives

33

Complete responses

45

Survey records

5

Countries represented

73.3%

Completion rate

THE CLINICAL QUESTION

When retinal AI and clinical judgment do not align, how are ophthalmologists preserving trust and accountability?

The dominant response is not to defer automatically to either side. Most respondents order additional imaging when an AI result conflicts with their interpretation, suggesting that disagreement is treated as a signal to investigate further.

Trust is also linked to how inspectable and operationally integrated the tool is. Explainability matters to most respondents, real-world image complexity remains a common failure source, and accountability is still placed mainly on the signing clinician or the organization deploying the system.

Coverage: United States, United Kingdom, Italy, France, Germany. Survey findings are presented as the opinions and practice patterns reported by participating professionals.

EXECUTIVE SUMMARY

The pattern in one view

The survey shows retinal AI gaining value through specific, verifiable tasks rather than broad diagnostic autonomy. The most important tensions appear at the boundaries: difficult images, conflicting interpretations, decentralized screening, workflow integration, and responsibility for harm.

01 / SIGNAL

OCT augmentation has the clearest utility

Advanced segmentation and quantification are far more commonly selected than predictive analytics as the most useful current AI function.

02 / SIGNAL

Disagreement prompts more evidence

Alternative imaging is the dominant response when human and AI interpretations conflict.

03 / SIGNAL

Complex eyes remain difficult for algorithms

Overlapping pathology and media artifacts lead the reported failure modes.

04 / SIGNAL

Responsibility remains close to care delivery

Clinicians and healthcare organizations carry most of the accountability in respondents' views.

Reading across the findings: The data separates technical complexity from clinical trust. AI is most valued where it performs visible measurement, while disagreement prompts more imaging, complex pathology remains a failure point, and accountability stays close to the clinician and organization. Explainability and workflow integration therefore are not cosmetic product features in this dataset; they are part of the conditions under which an algorithm becomes usable.

SECTION 01

Current value is concentrated in OCT analysis rather than prediction

CONTEXT

Retinal AI includes several very different kinds of tools. Some quantify anatomy already visible on OCT, while others screen autonomously or predict future disease. Those functions demand different levels of trust.

DATA

56.8% select advanced OCT segmentation as the AI application with the highest current clinical utility, compared with 21.6% for workflow triage, 16.9% for autonomous community screening, and 2.7% for predictive analytics.

INSIGHT

Utility is strongest where AI performs repeatable measurement on a scan that the ophthalmologist can still inspect directly.

WHY IT MATTERS

The result suggests that augmentation is maturing faster than prediction in the eyes of this cohort.

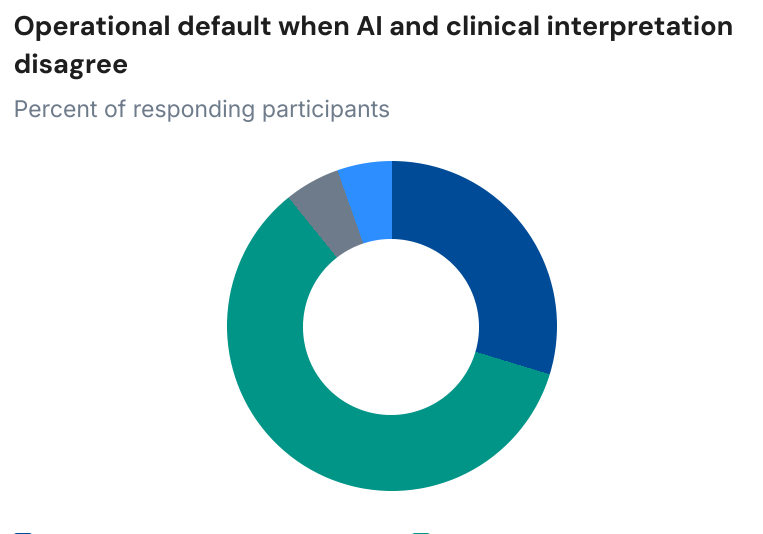
This makes intuitive clinical sense. Automated layer boundaries, fluid volumes, or atrophy area can reduce measurement burden while leaving the underlying image visible for verification. A predictive score, by contrast, asks the clinician to trust a future risk estimate whose causal path may be less obvious.

The survey therefore distinguishes between "AI in ophthalmology" as a broad concept and the specific tasks that have earned practical confidence.

The significance of this finding becomes clearer when it is read against the rest of the dataset. Ophthalmologists favor AI where it can augment a familiar imaging task with quantification, rather than where it asks them to act on a forecast about a future event. Layer segmentation, fluid measurement, and atrophy tracking produce visible outputs that can be checked against the raw scan, while prediction requires a different level of confidence in model generalization. The percentage therefore sits inside a broader decision environment rather than operating as a stand-alone trigger.

AI implementation providing the greatest current clinical utility

Percent of responding participants



CONTEXT

Human-AI disagreement is the critical stress test for a diagnostic tool. When both sources agree, workflow is easy. When they diverge, the clinic needs a rule for what happens next.

DATA

59.5% order alternative imaging, 29.7% default to their own clinical judgment, 5.4% lean toward AI, and 5.4% seek external peer review.

INSIGHT

The dominant operational default is evidence expansion rather than hierarchy.

WHY IT MATTERS

A disagreement can trigger OCT-A, fluorescein angiography, comparison with prior scans, or other assessment that reframes the problem with new information.

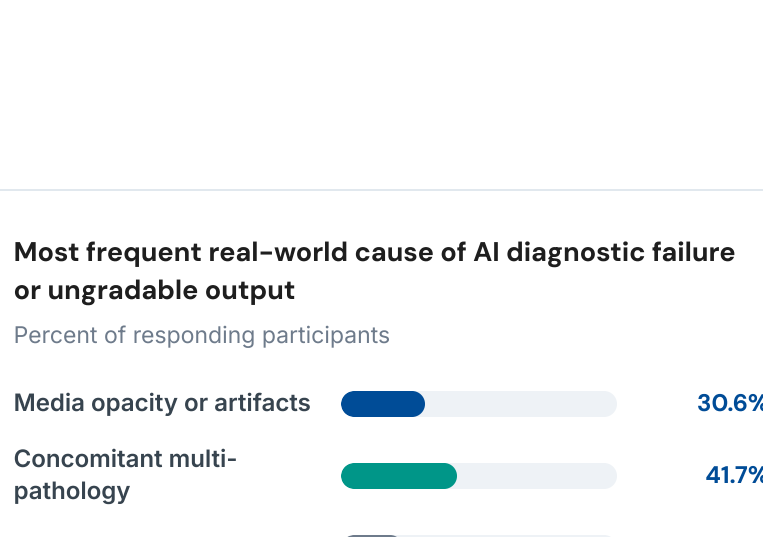
This behavior also limits classic automation bias. Only a small minority report leaning toward the AI finding when it conflicts with their own interpretation.

The finding suggests that ophthalmologists are incorporating AI into diagnostic reasoning without treating regulatory clearance as a substitute for case-specific judgment.

The significance of this finding becomes clearer when it is read against the rest of the dataset. Disagreement is being managed as an uncertainty problem. The dominant response is to generate additional evidence. Retinal imaging offers multiple complementary modalities, so OCT-A, fluorescein angiography, or other imaging can help resolve whether the discrepancy reflects missed pathology, image artifact, or model limitation. The percentage therefore sits inside a broader decision environment rather than operating as a stand-alone trigger.

Operational default when AI and clinical interpretation disagree

Percent of responding participants



SECTION 02

Human-AI disagreement usually triggers more evidence

CONTEXT

Human-AI disagreement is the critical stress test for a diagnostic tool. When both sources agree, workflow is easy. When they diverge, the clinic needs a rule for what happens next.

DATA

59.5% order alternative imaging, 29.7% default to their own clinical judgment, 5.4% lean toward AI, and 5.4% seek external peer review.

INSIGHT

The dominant operational default is evidence expansion rather than hierarchy.

WHY IT MATTERS

A disagreement can trigger OCT-A, fluorescein angiography, comparison with prior scans, or other assessment that reframes the problem with new information.

This behavior also limits classic automation bias. Only a small minority report leaning toward the AI finding when it conflicts with their own interpretation.

The finding suggests that ophthalmologists are incorporating AI into diagnostic reasoning without treating regulatory clearance as a substitute for case-specific judgment.

The significance of this finding becomes clearer when it is read against the rest of the dataset. Disagreement is being managed as an uncertainty problem. The dominant response is to generate additional evidence. Retinal imaging offers multiple complementary modalities, so OCT-A, fluorescein angiography, or other imaging can help resolve whether the discrepancy reflects missed pathology, image artifact, or model limitation. The percentage therefore sits inside a broader decision environment rather than operating as a stand-alone trigger.

Most frequent real-world cause of AI diagnostic failure or ungradable output

Percent of responding participants



SECTION 03

Multi-pathology is the most common reported AI failure condition

CONTEXT

Model performance is often described using clean validation datasets, but clinic images are affected by disease overlap, opacity, anatomy, and patient cooperation. These conditions test whether an algorithm generalizes beyond its ideal input.

DATA

41.7% identify concomitant multi-pathology as the leading real-world failure source, 30.6% media opacity or artifact, 22.2% anatomical variants, and 5.6% fixation or centering problems.

INSIGHT

Complex biology is a larger concern than simple acquisition failure in this cohort.

WHY IT MATTERS

A model that performs well on one disease phenotype may become less reliable when several pathologies alter the same retinal structures.

This is especially relevant in older or highly comorbid patients, where cataract, diabetic change, macular degeneration, membranes, myopia, and vascular disease can coexist.

The pattern also strengthens the case for explicit uncertainty outputs and robust "ungradable" pathways rather than forcing a binary answer from a model confronted with unfamiliar anatomy.

The significance of this finding becomes clearer when it is read against the rest of the dataset. The algorithm appears most vulnerable when the eye does not resemble a clean single-disease example. Complexity in the patient becomes complexity in the image. Severe myopia, membranes, diabetic changes, cataract, haze, and anatomical variation can alter features that models rely on, especially when training data underrepresent overlapping conditions. The percentage therefore sits inside a broader decision environment rather than operating as a stand-alone trigger.

Primary concern with decentralized autonomous retinal screening

Percent of responding participants



CONTEXT

Autonomous screening can extend retinal assessment beyond specialist clinics, but a disease-specific system changes who controls the referral loop and what a negative screen means to the patient.

DATA

33.3% are most concerned about missed incidental pathology, 30.6% false-positive referral influx, 19.4% fragmentation of care, and 16.7% false-negative delayed treatment.

INSIGHT

The main anxieties are system-level. They concern what happens around the algorithm, not only whether it detects its target disease.

WHY IT MATTERS

A technically successful screening program can still create clinical problems if the referral network, scope of screening, and patient messaging are unclear.

The FDA definition of retinal diagnostic software is tied to specific screening purposes. That regulatory specificity matters because a result for referable diabetic retinopathy is not the same as a comprehensive assessment for glaucoma, nevi, macular disease, or other incidental pathology.

The survey therefore points to a communication challenge: autonomous screening has to be framed precisely enough that a "negative" result is not misread as "the eye is healthy."

The significance of this finding becomes clearer when it is read against the rest of the dataset. The dominant concern is broader than diabetic retinopathy detection accuracy. It is the system effect created around a narrow screening target. A platform optimized for one retinal condition may not provide a comprehensive overview, yet patients or non-specialist providers may interpret a negative result more broadly than intended. The percentage therefore sits inside a broader decision environment rather than operating as a stand-alone trigger.

SECTION 04

Autonomous screening raises concern about what the algorithm is not looking for

CONTEXT

Autonomous screening can extend retinal assessment beyond specialist clinics, but a disease-specific system changes who controls the referral loop and what a negative screen means to the patient.

DATA

33.3% are most concerned about missed incidental pathology, 30.6% false-positive referral influx, 19.4% fragmentation of care, and 16.7% false-negative delayed treatment.

INSIGHT

The main anxieties are system-level. They concern what happens around the algorithm, not only whether it detects its target disease.

WHY IT MATTERS

A technically successful screening program can still create clinical problems if the referral network, scope of screening, and patient messaging are unclear.

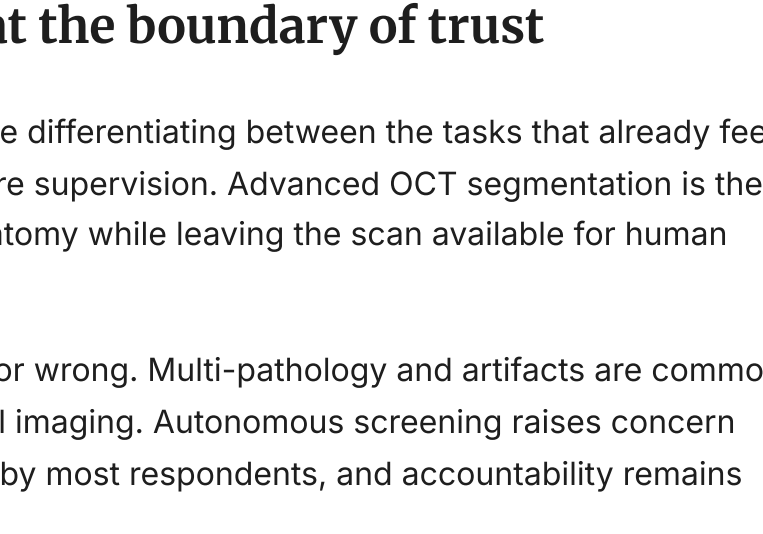
The FDA definition of retinal diagnostic software is tied to specific screening purposes. That regulatory specificity matters because a result for referable diabetic retinopathy is not the same as a comprehensive assessment for glaucoma, nevi, macular disease, or other incidental pathology.

The survey therefore points to a communication challenge: autonomous screening has to be framed precisely enough that a "negative" result is not misread as "the eye is healthy."

The significance of this finding becomes clearer when it is read against the rest of the dataset. The dominant concern is broader than diabetic retinopathy detection accuracy. It is the system effect created around a narrow screening target. A platform optimized for one retinal condition may not provide a comprehensive overview, yet patients or non-specialist providers may interpret a negative result more broadly than intended. The percentage therefore sits inside a broader decision environment rather than operating as a stand-alone trigger.

Where primary accountability belongs after an AI miss causing preventable vision loss

Percent of responding participants



CONTEXT

Responsibility becomes more complicated as software autonomy increases. The survey asks where accountability belongs when a clear vision-threatening lesion is missed and preventable harm follows.

DATA

37.1% place primary responsibility on the signing clinician, 34.3% on the healthcare organization, 20.0% favor dynamic shared liability, and 8.6% place it on the developer or vendor.

INSIGHT

Most respondents anchor accountability close to the clinical deployment environment.

WHY IT MATTERS

If the clinician and organization expect to carry risk, they are likely to demand strong validation, clear override pathways, documentation and workflow controls.

This helps explain why reimbursement and EHR integration matter as well. A clinic may be reluctant to accept additional review burden and liability if the tool also creates extra clicks or lacks a sustainable payment model.

The accountability pattern is therefore not separate from adoption. It is one of the factors that determines how much autonomy clinicians are willing to give the system.

The significance of this finding becomes clearer when it is read against the rest of the dataset. Even in an AI-enabled workflow, respondents locate responsibility mainly where the response is selected, deployed, interpreted, and signed off. Clinical decisions occur inside organizational systems. The model can generate an output, but clinicians and institutions still control how it enters the chart, how disagreements are handled, and what escalation occurs. The percentage therefore sits inside a broader decision environment rather than operating as a stand-alone trigger.

SECTION 05

Accountability remains anchored to human and institutional actors

CONTEXT

Responsibility becomes more complicated as software autonomy increases. The survey asks where accountability belongs when a clear vision-threatening lesion is missed and preventable harm follows.

DATA

37.1% place primary responsibility on the signing clinician, 34.3% on the healthcare organization, 20.0% favor dynamic shared liability, and 8.6% place it on the developer or vendor.

INSIGHT

Most respondents anchor accountability close to the clinical deployment environment.

WHY IT MATTERS

If the clinician and organization expect to carry risk, they are likely to demand strong validation, clear override pathways, documentation and workflow controls.

This helps explain why reimbursement and EHR integration matter as well. A clinic may be reluctant to accept additional review burden and liability if the tool also creates extra clicks or lacks a sustainable payment model.

The accountability pattern is therefore not separate from adoption. It is one of the factors that determines how much autonomy clinicians are willing to give the system.

The significance of this finding becomes clearer when it is read against the rest of the dataset. Even in an AI-enabled workflow, respondents locate responsibility mainly where the response is selected, deployed, interpreted, and signed off. Clinical decisions occur inside organizational systems. The model can generate an output, but clinicians and institutions still control how it enters the chart, how disagreements are handled, and what escalation occurs. The percentage therefore sits inside a broader decision environment rather than operating as a stand-alone trigger.

SECTION 06

Explainability is reassurance for most, but a hard requirement for a minority

CONTEXT

Explainability is often presented as a technical feature, but in clinical imaging it can function as a communication layer between algorithm and clinician.

DATA

25.7% describe explainability as mandatory, 42.9% as secondary reassurance, 14.3% give it low importance, and 17.1% consider it irrelevant because they perform a full manual review.

INSIGHT

Most respondents value some ability to inspect why the model reached its conclusion, especially when the case is borderline.

WHY IT MATTERS

Heatmaps, confidence estimates, and lesion annotations can help a clinician decide whether the output is anatomically plausible or whether further review is needed.

Explainability does not guarantee correctness. A persuasive heatmap can still be wrong. Its value is that it exposes part of the model's reasoning process to clinical scrutiny rather than leaving the output as an unexplained label.

The survey suggests that this inspectability may be an important adoption feature even when the clinician remains responsible for reviewing the raw scan.

The significance of this finding becomes clearer when it is read against the rest of the dataset. Trust is being built through inspectability rather than statistical performance alone, especially in borderline cases. Visual explanations give clinicians a way to compare the model's attention with the raw anatomy and their own interpretation, making disagreement easier to interrogate. The percentage therefore sits inside a broader decision environment rather than operating as a stand-alone trigger.

Importance of explainability features to trust in AI output

Percent of responding participants

Explainability does not guarantee correctness. A persuasive heatmap can still be wrong. Its value is that it exposes part of the model's reasoning process to clinical scrutiny rather than leaving the output as an unexplained label.

The survey suggests that this inspectability may be an important adoption feature even when the clinician remains responsible for reviewing the raw scan.

The significance of this finding becomes clearer when it is read against the rest of the dataset. Trust is being built through inspectability rather than statistical performance alone, especially in borderline cases. Visual explanations give clinicians a way to compare the model's attention with the raw anatomy and their own interpretation, making disagreement easier to interrogate. The percentage therefore sits inside a broader decision environment rather than operating as a stand-alone trigger.

The likely near-term model is therefore supervised intelligence: software performs measurement, prioritization, and selected screening tasks, while ophthalmologists retain control of ambiguity, cross-pathology interpretation, escalation, and the final clinical meaning of the image.

Questions professionals continue to ask

This topic remains dynamic. The questions below reflect common practical searches around the field rather than retrospective questions about this survey.

How is AI used in retinal imaging today?

Current applications include diabetic retinopathy screening, OCT segmentation, fluid and atrophy quantification, scan triage, and research into predictive models for disease progression.

Can AI diagnose retinal disease without an ophthalmologist?

Some autonomous systems are cleared for specific screening tasks, such as referable diabetic retinopathy. Their intended use is narrower than a comprehensive ophthalmic examination, and deployment depends on the device and local care pathway.

Why can retinal AI fail on real-world scans?

Image artifacts, media opacity, overlapping diseases, anatomical variation, poor fixation, and populations that differ from model training data can all reduce performance or produce ungradable results.

What happens when an ophthalmologist disagrees with an AI result?

The response depends on the clinical situation. Additional imaging, closer manual review, comparison with prior scans, or subspecialty review can help resolve uncertainty.

What does explainable AI mean in ophthalmology?

Explainability features can include heatmaps, confidence information, lesion boundaries, or other visual cues that help clinicians understand what image features contributed to the model output.

Who is responsible when clinical AI makes an error?

Accountability depends on law, regulation, device autonomy, organizational policy, and how the clinician uses the output. The issue remains an active area of medical, legal, and regulatory development.

Explore the complete insight. Contribute to what comes next.

MDForLives turns real healthcare perspectives into research-driven insights that reveal patterns in care, decisions, and lived experience.

Join the MDForLives Panel · Shape the next healthcare insight report

Join the Panel